The evolution of interdependence in a four-way mealybug symbiosis

The Evolution of Interdependence in a Four-Way Mealybug Symbiosis

Arkadiy I. Garber[1,2], Maria Kupper[1,2], Dominik R. Laetsch[3], Stephanie R. Weldon[1], Mark S. Ladinsky[4], Pamela J. Bjorkman[4], and John P. McCutcheon[1,2,*]

[1]Division of Biological Sciences, University of Montana, Missoula, Montana, USA

[2]Biodesign Center for Mechanisms of Evolution and School of Life Sciences, Arizona State University, Tempe, Arizona, USA

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

[4]Department of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA

*Corresponding author: E-mail: john.mccutcheon@asu.edu.

Accepted: 24 May 2021

Abstract

Mealybugs are insects that maintain intracellular bacterial symbionts to supplement their nutrient-poor plant sap diets. Some mealybugs have a single betaproteobacterial endosymbiont, a *Candidatus* Tremblaya species (hereafter *Tremblaya*) that alone provides the insect with its required nutrients. Other mealybugs have two nutritional endosymbionts that together provision these same nutrients, where *Tremblaya* has gained a gammaproteobacterial partner that resides in its cytoplasm. Previous work had established that *Pseudococcus longispinus* mealybugs maintain not one but two species of gammaproteobacterial endosymbionts along with *Tremblaya*. Preliminary genomic analyses suggested that these two gammaproteobacterial endosymbionts have large genomes with features consistent with a relatively recent origin as insect endosymbionts, but the patterns of genomic complementarity between members of the symbiosis and their relative cellular locations were unknown. Here, using long-read sequencing and various types of microscopy, we show that the two gammaproteobacterial symbionts of *P. longispinus* are mixed together within *Tremblaya* cells, and that their genomes are somewhat reduced in size compared with their closest nonendosymbiotic relatives. Both gammaproteobacterial genomes contain thousands of pseudogenes, consistent with a relatively recent shift from a free-living to an endosymbiotic lifestyle. Biosynthetic pathways of key metabolites are partitioned in complex interdependent patterns among the two gammaproteobacterial genomes, the *Tremblaya* genome, and horizontally acquired bacterial genes that are encoded on the mealybug nuclear genome. Although these two gammaproteobacterial endosymbionts have been acquired recently in evolutionary time, they have already evolved codependencies with each other, *Tremblaya*, and their insect host.

Key words: pseudogenes, endosymbionts, mealybugs, genome reduction, transposases, metabolic interdependence.

Significance

Mealybugs are sap-feeding insects that house between one and three bacterial endosymbionts to supplement their nutritionally poor diets. Many mealybug–bacteria relationships were established tens or hundreds of millions of years ago, and these ancient examples show high levels host-endosymbiont genomic and metabolic integration. Here, we describe the complete genomes and cellular locations for two bacterial endosymbionts which have recently transitioned from a free-living to an intracellular state. Our work reveals the rapid emergence of metabolic interdependence between these two nascent endosymbionts, their partner bacterial cosymbiont in whose cytoplasm they reside, and their insect host cell. Our work reaffirms that intracellular bacteria rapidly adapt to a host-restricted lifestyle through breakage or loss of redundant genes.

Introduction

Insects with nutrient-poor diets (e.g., plant sap, blood, wood) maintain microbial symbionts that supplement their diet with compounds such as amino acids and vitamins (Baumann 2005; Douglas 2006). Mealybugs (fig. 1*A*) are insects that exclusively consume phloem sap and maintain nutritional endosymbiotic bacteria within specialized cells called bacteriocytes (Buchner 1965; von Dohlen et al. 2001; Baumann et al. 2002). Mealybug bacteriocytes house between one and three different bacterial endosymbionts depending on the mealybug species (Kono et al. 2008; Koga et al. 2013; López-Madrigal et al. 2013; Husník and McCutcheon 2016; Szabó et al. 2017; Gil et al. 2018). These mealybug endosymbionts produce essential amino acids and vitamins, which are present in plant sap at levels insufficient for insect growth. Although it is not uncommon for insects to simultaneously maintain multiple endosymbionts (Buchner 1965; Fukatsu and Nikoh 1998; Thao et al. 2002; Toh et al. 2006; Moran et al. 2008; McCutcheon and Moran 2010), the spatial organization of the dual mealybug endosymbiosis is unusual: each bacteriocyte house cells of *Candidatus* Tremblaya princeps (betaproteobacteria, hereafter referred to as *Tremblaya*), and inside each *Tremblaya* cell reside tens to hundreds of cells of another endosymbiont from the gammaproteobacterialfamily Enterobacteriaceae (von Dohlen et al. 2001; Downie and Gullan 2005; Gatehouse et al. 2012), with the titer of gammaproteobacterial symbionts varying depending on host species and developmental stage (Kono et al. 2008; Parkinson et al. 2017). Many of these intra-*Tremblaya* endosymbionts are members of the *Sodalis* genus, which are well-represented among endosymbionts of insects (Toh et al. 2006; Clayton et al. 2012; Oakeson et al. 2014; Husník and McCutcheon 2016; McCutcheon et al. 2019; Hall et al. 2020).

Genomic studies of numerous insect–endosymbiont systems have revealed strong and consistent patterns of complementary gene loss and retention among all members of the symbiosis (Shigenobu et al. 2000; van Ham et al. 2003; Wu et al. 2006; McCutcheon and Moran 2010; Sloan and Moran 2012; Łukasik et al. 2018). Although, in most cases, a single endosymbiont genome will retain complete or near-complete pathways for individual metabolites, mealybug endosymbionts are unusual in that the reciprocal pattern of gene loss and retention exists within biochemical pathways (McCutcheon and von Dohlen 2011; López-Madrigal et al. 2013; Husník and McCutcheon 2016; Szabó et al. 2017; Gil et al. 2018). Most of the previously published mealybug endosymbiont genomes were highly reduced in size (less than 1 Mb) and gene dense (containing few pseudogenes), which made discerning these complementary gene loss and retention patterns relatively straightforward.

Pseudococcus longispinus harbors the symbiont *Tremblaya*, but unlike *Tremblaya* in other mealybugs, the *P. longispinus* strain of *Tremblaya* house not one but two gammaproteobacterial endosymbionts (Rosenblueth et al. 2012; Husník and McCutcheon 2016). We previously reported draft genome assemblies of these two gammaproteobacterial endosymbionts, which suggested that their combined genome sizes were large, approximately 8.2 megabase pairs (Mbp) in length (Husník and McCutcheon 2016). Phylogenetic analysis showed that one of these gammaproteobacterial symbionts belonged to the *Sodalis* genus, and the other was more closely related to members of the *Pectobacterium* genus. However, the poor quality of these draft genome assemblies made detailed genomic analysis impossible. Light microscopy on *P. longispinus* (Gatehouse et al. 2012) suggested that the gammaproteobacterial endosymbionts reside inside *Tremblaya* cells, as is the case in other mealybugs (von Dohlen et al. 2001). But it was unclear from these data whether one or both of these gammaproteobacteria are restricted to *Tremblaya* cells (i.e., if they are also in the cytoplasm of the host insect bacteriocyte), whether each gammaproteobacterial species is restricted to particular *Tremblaya* cell types, or whether the two gammaproteobacterial symbionts are mixed together inside undifferentiated *Tremblaya* cells. Here, we add long-read data generated from *P. longispinus* bacteriome tissue to greatly improve the gammaproteobacterial genome assemblies and annotations. We describe the relative cellular locations of the endosymbionts using fluorescence and transmission electron microscopy and report the genome evolutionary patterns and metabolic contributions of the microbial members of this unusual four-way symbiosis.

Results

Gammaproteobacterial Endosymbionts Are Located within *Tremblaya*

Previous light microscopy suggested that at least some, if not all, of the gammaproteobacterial endosymbionts of *P. longispinus* resided inside of *Tremblaya* (Gatehouse *et al.* 2012). However, these data lacked the resolution to clarify whether or not both species of gammaproteobacterial endosymbionts were exclusively contained within *Tremblaya* or whether some might also live in the cytoplasm of bacteriocytes. We used transmission electron microscopy (TEM) to identify the localization of the gammaproteobacterial cells. Our TEM data suggest that all gammaproteobacterial endosymbiont cells are contained within *Tremblaya* and are not free in the cytoplasm of the host insect cell (fig. 1*B–E*). At low magnification, elongated cells of different shapes and sizes, which we presume to be the gammaproteobacterial symbionts, can be seen inside of *Tremblaya* cells (fig. 1*B*). Structures of roughly similar size and electron density can be observed outside of *Tremblaya* cells, but these were found to be mitochondria upon examination at higher magnification (fig. 1*C* and *D*). We note

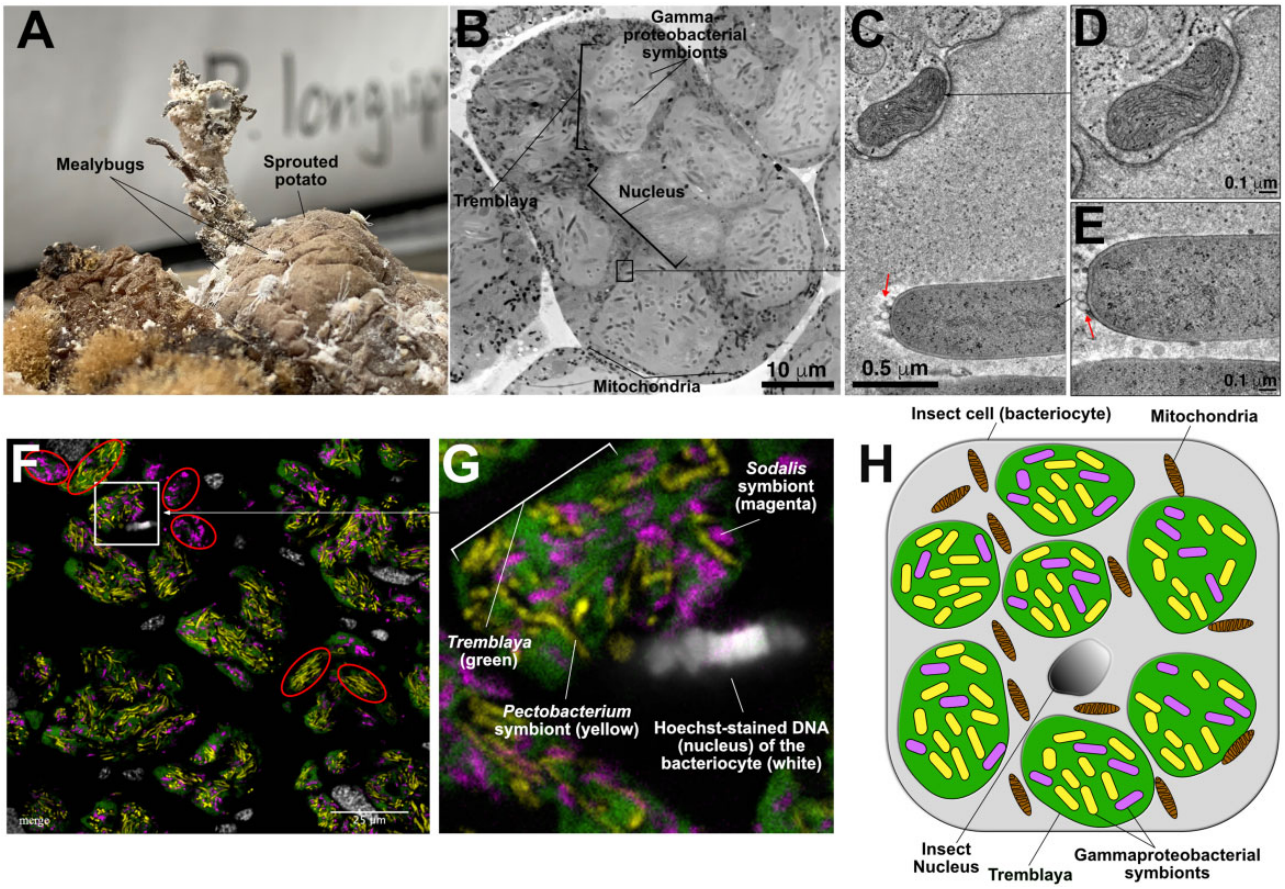

Fig. 1.—The structure of the *Pseudococcus longispinus* symbiosis. (*A*) Image of *P. longispinus* mealybugs on a sprouted potato. (*B*) Montaged TEM overview image of a bacteriocyte from *P. longispinus*. The six to seven light gray blobs are *Tremblaya* cells, surrounding a central eukaryotic nucleus. Within each *Tremblaya* cell reside rod-shaped and more electron-dense gammaproteobacterial cells. Black-colored rods in between *Tremblaya* are mitochondria within eukaryotic cytoplasm. The insect nucleus is at the center of the bacteriocyte in a gray shade that is similar to *Tremblaya*. (*C*) Details from an electron tomographic slice showing the boundary of a *Tremblaya* cell, where a mitochondrion is visible near the *Tremblaya* cell envelope. (*D*) Higher magnification view of the mitochondrion shown in *C*. (*E*) Tomographic slice of a gammaproteobacterial symbiont that resides inside *Tremblaya*, showing numerous outer membrane vesicles (red arrows). The bacterial symbionts are easily distinguished from eukaryotic mitochondria. (*F*) Fluorescent in situ hybridization (FISH) image of *P. longispinus* bacteriome tissue showing the localization of two different gammaproteobacterial endosymbionts within *Tremblaya* cells. Fluorophore-labeled probes were used to localize *Tremblaya* cells (green) and the two gammaproteobacterial endosymbionts (yellow and magenta). Tremblaya cells that appear to harbor exclusively, or almost-exclusively, one type of gammaproteobacterial endosymbiont are circled in red. DNA and, therefore, insect nuclei were counterstained with Hoechst (white). Each nucleus is surrounded by several *Tremblaya* cells per bacteriocyte. (*G*) Zoomed in and annotated detail fluorescence microscopy image of a *P. longispinus* bacteriocyte. (*H*) Schematic representation of *P. longispinus* bacteriocytes.

numerous outer membrane vesicles (OMVs, Toyofuku et al. 2019) apparently being extruded by the *Sodalis*- or *Pectobacterium*-related endosymbiont cells (fig. 1*E*). The function of these OMVs in the symbiosis, if any, is unknown.

Using previously reported small subunit (SSU) ribosomal RNA (rRNA) sequences (Husník and McCutcheon 2016), we next performed fluorescence in situ hybridization (FISH) targeting the SSU rRNA to establish the relative locations of the two gammaproteobacterial endosymbionts. Here, we were testing whether there were two different types of *Tremblaya* cells, each containing only one type of gammaproteobacterial cell, or whether both gammaproteobacterial species were mixed together inside of one type of *Tremblaya* cell. We find that both gammaproteobacterial endosymbionts are mixed together in one type of *Tremblaya* cell (fig. 1*F*–*H*). The overall distribution of endosymbionts within *Tremblaya* cells suggests that the *Pectobacterium* relative is more abundant than the *Sodalis* relative (colored yellow and violet, respectively, in fig. 1*F* and *G*). We note that there are some *Tremblaya* cells or regions of *Tremblaya* cells where the *Pectobacterium* relative appears highly abundant, with almost no cells of the *Sodalis* relative present, and vice versa, with some *Tremblaya* regions containing mostly cells of the *Sodalis*

relative (red circles in fig. 1*F*). It is important to note that the image in figure 1*F* is a single plane of focus in a Z-stack; thus, the apparent absence of one of the endosymbionts from a *Tremblaya* cell does not necessarily indicate that endosymbiont is definitively missing there. Cells of the *Pectobacterium* relative appear to be longer than cells of the *Sodalis* relative. This mixture of long and short cells is consistent with what we see in the TEM images, although the identities of the two cell types cannot be discerned in TEM (fig. 1*B*).

Gammaproteobacterial Endosymbionts Have Large Genomes Similar to Free-Living Bacteria

Previous efforts to assemble the genomes of the two gammaproteobacterial symbionts using only short-read Illumina technology resulted in highly fragmented genome assemblies (Husník and McCutcheon 2016). Our addition of PacBio reads greatly improved the quality of the *Sodalis* symbiont's genome (3 vs. ~200 contigs from short reads alone) due to their ability to span repetitive insertion sequences (ISs) that appear to be abundant in that genome (supplemental file 1, Supplementary Material online). The *Pectobacterium*-related symbiont genome was also improved by long-read sequencing (9 vs. 40 contigs from short reads alone). Despite numerous computational and polymerase chain reaction-based experiments, we were unable to close either genome into complete circular-mapping molecules. To aid in the binning of contigs for each endosymbiont genome, we relied on their differential read coverage calculated from Illumina short reads as well as the similarity of endosymbiont genes compared with their nonendosymbiont *Sodalis* and *Pectobacterium* relatives. At 4.3 and 3.6 Mb, both gammaproteobacterial endosymbiont genomes are similar in size to genomes of many free-living bacteria. One circular-mapping contig of approximately 150 kb was putatively assigned as a plasmid to the *Pectobacterium* relative, based on its circular structure and its predicted genes showing high similarity to genes from other *Pectobacterium* and *Brenneria* spp. One additional contig of approximately 90 kb, containing genes with high similarity to other *Sodalis* spp. genes, showed 3-fold higher coverage relative to the other *Sodalis*-related contigs; this could be a plasmid or a large repeat region of the genome. Finally, we identified an *Arsenophonus*-related plasmid, encoding mostly hypothetical proteins, in addition to eight genes annotated as Type II and IV secretion proteins, although 71 of the total 101 predicted genes on this plasmid appear to be pseudogenized. As no other *Arsenophonus* contigs are found in our assembly, we assume that this plasmid resides in one of the two gammaproteobacteria, although we did not pursue the cellular location of this plasmid further. An overview of the endosymbiont genomes is shown in table 1.

Mapping of Illumina sequence reads to the endosymbiont genomes indicates that the *Pectobacterium* endosymbiont is twice as abundant as the *Sodalis* endosymbiont. This is consistent with FISH images, where the *Pectobacterium*-related cells appear more abundant than the *Sodalis*-related cells (fig. 1*B*). We note that using read mapping as a proxy for endosymbiont abundance may be prone to error because endosymbiont genomes are known to exist in multiple copies per cell. For example, Illumina read mapping to the genome of *Tremblaya* suggests that it is approximately 25 times more abundant than its gammaproteobacterial cosymbionts, but *Tremblaya* cells are clearly less abundant than gammaproteobacterial cells because each *Tremblaya* cell contains numerous gammaproteobacterial cells (fig. 1). It is likely that *Tremblaya*'s genome is present in hundreds or thousands of copies per cell, consistent with previous reports of extreme polyploidy in ancient endosymbionts with tiny genomes, such as *Candidatus* Hodgkinia cicadicola (Van Leuven et al. 2014), *Candidatus* Sulcia muelleri (Woyke et al. 2010), and *Buchnera aphidicola* (Komaki and Ishikawa 1999).

Gammaproteobacterial Symbionts Are Related to Opportunistic Pathogens Known to Infect Insects

The closest sequenced nonendosymbiont relatives of the *P. longispinus* gammaproteobacterial symbionts are *Pectobacterium wasabiae* (average amino acid identity = 76.1%) and *Sodalis praecaptivus* HS (average amino acid identity = 86.0%; hereafter, *Sodalis* HS). *Sodalis* HS was isolated from a human infection (Clayton et al. 2012; Chari et al. 2015), and its genome suggests that this bacterium may be an opportunistic pathogen capable of infecting animal and plant cells (Clayton et al. 2012). *P. wasabiae* is a known pathogen of plants and has been identified as the causative agent of potato soft rot (Gardan et al. 2003; Pasanen et al. 2013; Yuan et al. 2014). Phylogenomic analysis, using a concatenated set of 172 single-copy genes common to Gammaproteobacteria (Lee 2019), confirms the affiliation of one endosymbiont squarely within the *Sodalis* genus, closely related to other recently established endosymbionts such as *Ca*. S. glossinidius (hereafter, *S. glossinidius*) and *S. pierantonius* str. SOPE (hereafter, SOPE) (Husník and McCutcheon 2016) (fig. 2). Of note, SOPE was estimated to have been established as an insect endosymbiont from a *Sodalis* HS relative very recently, approximately 28,000 years ago (Clayton et al. 2012). Using the GC-content among 4-fold degenerate sites in the *Sodalis* endosymbiont of *P. longispinus* (in comparison with SOPE), and assuming a clock-like reduction in GC-content following host restriction, we estimate that its divergence from *Sodalis* HS occurred approximately 67,000 years ago, although we stress that this is a very rough estimate (supplementary file 2, Supplementary Material online). For example, selective pressures are likely different between the intra-*Tremblaya* space inside mealybugs and the insect intracytoplasmic space of weevils, where SOPE resides.

Table 1.
Assembly Summary for the Three Endosymbionts and Their Putative Plasmids Assembled from *Pectobacterium longispinus* Bacteriomes

Endosymbionts and Plasmids	Genome/Plasmid Size (bp)	Average Read Depth (Illumina)	Number of Contigs	Candidate Pseudogenes/ Intact Genes
Pectobacterium symbionts	4,343,494	43.08	9	3,093/2,814
Pectobacterium plasmid	148,954	47.04	1	126/900
Sodalis symbiont	3,638,256	22.33	3	1,887/2,405
Sodalis (possible plasmid)	89,872	56.70	1	71/48
Tremblaya[a]	144,042	1,565.37	1	11/123
Arsenophonus plasmid (unbinned)	64,583	160.98	1	71/101

Note.—The *Pectobacterium* and *Sodalis* symbionts are named *Symbiopectobacterium endolongispinus* and *Sodalis endolongispinus*, respectively (see Naming of the Two Gamma-proteobacterial Symbionts).

[a]Genome of *Tremblaya* taken from Husník and McCutcheon (2016).

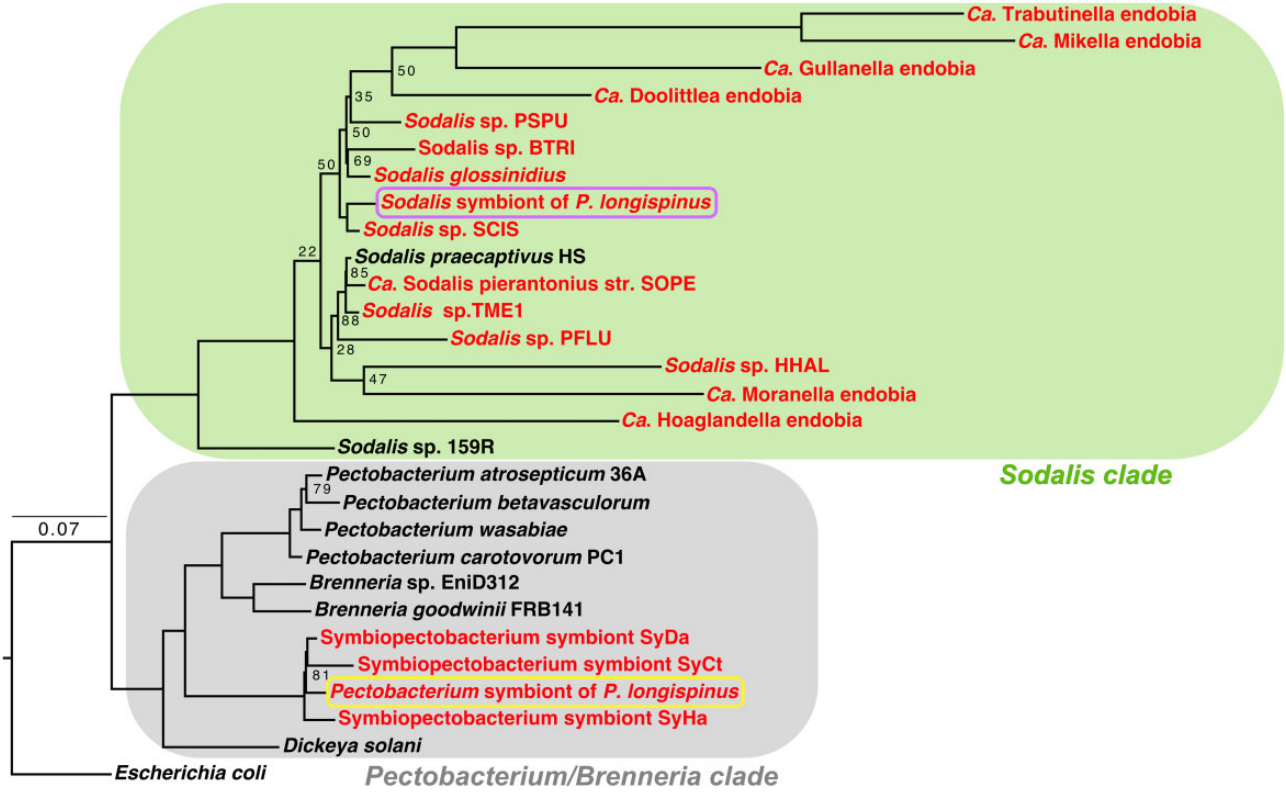

Fig. 2.—The gammaproteobacterial endosymbionts belong to two different groups. A phylogenomic tree constructed with a concatenated set of 172 single-copy genes designed for gammaproteobacteria (Lee 2019), of *Sodalis*- and *Pectobacterium-related* endosymbionts (colored red) and the closest free-living relatives (colored black). *Escherichia coli* genome is used as the outgroup. This tree reveals two distinct clades: one containing the *Pectobacterium-/Brenneria*-related bacteria and one containing the *Sodalis*-related bacteria. The two endosymbionts residing within *P. longispinus* bacteriocytes are emphasized in yellow (*Pectobacterium* related) and violet (*Sodalis* related) boxes. Unlabeled nodes have bootstrap support values greater than 90%.

The second gammaproteobacterial endosymbiont in *P. longispinus* is affiliated with *Pectobacterium* and *Brenneria* spp. and appears to fall within a newly proposed group of nematode and insect endosymbionts, named *Symbiopectobacterium* (Martinson et al. 2020). Blast-based comparison of open-reading frames confirms that these *Symbiopectobacterium*-clade symbionts are very closely related, sharing 94–97% average nucleic acid identity across their genomes (supplementary fig. 1, Supplementary Material online).

Naming of the Two Gammaproteobacterial Endosymbionts

For the *Sodalis* relative, we propose the name *Candidatus* Sodalis endolongispinus (hereafter, *Sod. endolongispinus*).

This name highlights its close phylogenetic relationship with other bacteria in the *Sodalis* genus (fig. 2) and its localization inside *P. longispinus* bacteriomes. Similarly, we propose the name *Candidatus* Symbiopectobacterium endolongispinus (hereafter, *Sym. endolongispinus*) for the *Pectobacterium* relative, reflecting its close phylogenetic relationship with the new Symbiopectobacterium group (Martinson et al. 2020), along with its localization inside *P. longispinus* bacteriomes.

Pseudogenes Abound in the Gammaproteobacterial Endosymbiont Genomes

Newly established endosymbionts contain unusually high numbers of pseudogenes compared with most bacterial genomes (Toh et al. 2006; Burke and Moran 2011; McCutcheon and Moran 2011; Clayton et al. 2012; Oakeson et al. 2014). Pseudogenes are thought to form as a bacterium transitions to a strict intracellular lifecycle because many previously essential genes are no longer required in the intracellular environment (Toh et al. 2006; Burke and Moran 2011; McCutcheon and Moran 2011). Additionally, rapid pseudogenization of some genes coding for immune-stimulating compounds, such as lipopolysaccharide, is likely to be adaptive for bacteria that have recently transitioned to an intracellular lifestyle (D'Souza and Kost 2016; McCutcheon et al. 2019).

The genomes of both gammaproteobacterial endosymbionts of *P. longispinus* contain thousands of pseudogenes (fig. 3 and supplementary table 1 and supplementary files 3, 4, Supplementary Material online). The coding densities of both of these genomes are approximately 50%, much lower than average for most other free-living bacteria (Ochman and Davalos 2006). Pseudogenes in *Sod. endolongispinus* and *Sym. endolongispinus* are found in nearly all gene categories, including membrane transport, amino acid metabolism, energy generation, secretion systems, transcriptional regulation, and motility. Several regions that appear to be remnants of prophages are also largely pseudogenized. Pseudogenes have been formed in a variety of ways, and some genes show multiple signs of pseudogenization (e.g., truncations and d*N*/d*S* values >0.3, Oakeson et al. 2014) (fig. 3*B*). A substantial proportion of pseudogenes were formed by frameshifts and nonsense mutations, resulting in partial gene deletions or multiple gene fragments derived from what used to be a single ancestral gene. Consequently, many predicted pseudogenes are shorter (labeled "truncated" in fig. 3*B*) relative to their closest, presumably functional, homologs in non-endosymbiotic bacteria (fig. 3*C*). Additionally, a small proportion of genes appear to be longer than their closest orthologs (labeled "run-on" in fig. 3*B*) likely due to a frameshift that results in the loss of a stop codon. Many putative pseudogenes or pseudogene fragments were unrecognizable to the prokaryotic gene-finding program Prodigal (Hyatt et al. 2010), possibly due to missing start/stop codons and/or frameshift-causing indels, and were only identified by performing BlastX (Camacho et al. 2009) searches of intergenic regions against NCBI's RefSeq database. We also detected cryptic pseudogenes, or genes that are structurally intact but are likely experiencing relaxed purifying selection, inferred from d*N*/d*S* ratios greater than 0.3 (Oakeson et al. 2014). However, we find that only a small proportion of predicted genes have elevated d*N*/d*S* values (0.5% of genes in *Sym. endolongispinus* and 2.2% in *Sod. endolongispinus*), suggesting that most genes are still experiencing strong purifying selection (fig. 3*C*).

Transposases Recently Proliferated within the Genome of Sod. endolongispinus

Sym. endolongispinus and *Sod. endolongispinus* were screened for ISs, which are types of mobile genetic elements in bacteria. ISs are typically made up of transposase genes along with other accessory and passenger genes (Mahillon and Chandler 1998) and have previously been suggested to proliferate during the early stages of host restriction in endosymbionts (Gil et al. 2008; Plague et al. 2008; Belda et al. 2010; Schmitz-Esser et al. 2011; Clayton et al. 2012; Oakeson et al. 2014). *Sod. endolongispinus* encodes at least 220 transposase genes, 96% of which are part of the IS3 family (supplementary fig. 2*A*, Supplementary Material online). The rest of the transposases are part of the ISNCY transposase family. Both of these IS families are also found in the close relative *Sodalis* HS, but in smaller numbers and different proportions. The expansion of IS3 family transposases in *Sod. endolongispinus* appears to have occurred very recently, because the vast majority of these transposases are part of two distinct clusters of paralogs, where each cluster contains approximately 80 nearly identical copies of the same transposase that has proliferated throughout the genome (supplementary file 1 and supplemental fig. 3, Supplementary Material online). Only a handful of transposase duplications have a d*S* value (a proxy for evolutionary divergence) greater than 0.5, which is the average d*S* of homologs between *Sod. endolongispinus* and *Sodalis* HS, suggesting that most transposition events occurred after divergence of the two species. Bouts of recent transposition are further supported by phylogenetics, where transposase sequences from *Sod. endolongispinus* and *Sym. endolongispinus* group together in highly similar clades to the exclusion of *Sodalis* HS and *P. wasabiae* transposases (supplementary fig. 4*A*, Supplementary Material online). There is one cluster of transposases, where single *Sod. endolongispinus* and *Sym. endolongispinus* sequences group with single *Sodalis* HS and *P. wasabiae* sequences, suggesting that at least in some cases, transposase duplications (and possibly horizontal gene transfers [HGT] events) might have occurred prior to speciation (supplementary fig. 4*B*, Supplementary Material online).

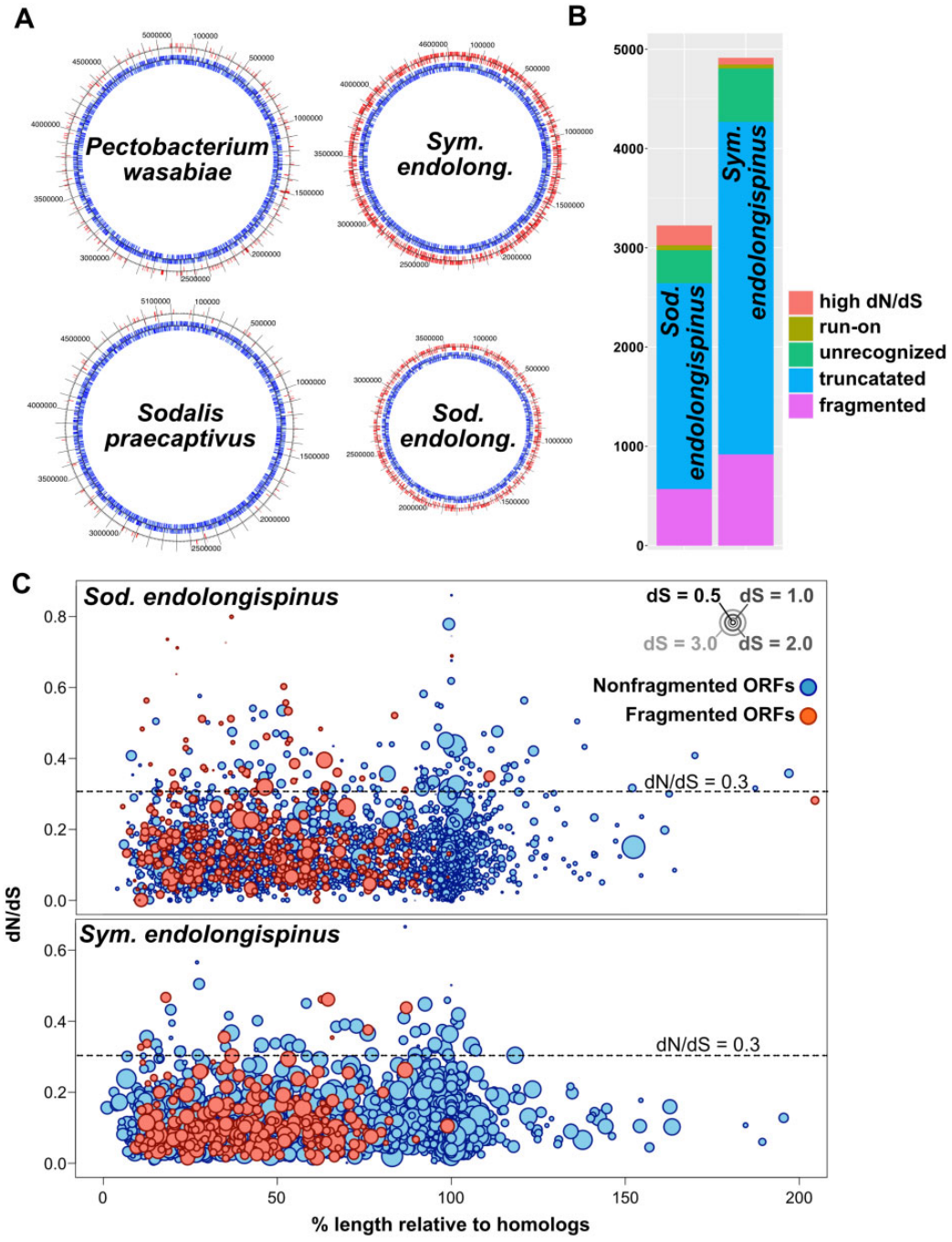

Fig. 3.—The features of gammaproteobacterial pseudogenes. (*A*) Genome maps showing the positions of candidate pseudogenes in the two *Pectobacterium longispinus* gammaproteobacterial endosymbiont genomes and their closest free-living relatives. Input genes (predicted using Prokka/Prodigal) are on the inner tracks and colored blue. Predicted pseudogenes are on the outer tracks and colored red. Numbers next to tick marks indicate the genome position (in bp) of each respective tick mark. (*B*) Summary of the types of gene disruptions occurring in each of the gammaproteobacterial symbionts. The total number of disruptions is greater than the total number of pseudogenes in each genome because many pseudogenes have more than one type of disruption. Truncated and run-on genes are those that are shorter or longer, respectively, relative to closest orthologs in NCBI's nr database; unrecognized genes are those that were failed to be identified by Prodigal's gene-finding algorithm but recruited more than five orthologs from NCBI's reference database; fragmented genes are those that are present in more than one ORF but appear to be derived from a single ancestral gene. (*C*) Plots showing gene degradation of endosymbiont genes. Each circle represents an endosymbiont gene; the *x*-axis represents the length of each gene relative to its ortholog in the reference genome (*Sodalis* HS or *P. wasabiae*); the *y*-axis represents d*N*/d*S* of each gene relative to its ortholog in the reference genome; genes that have truncating stop codons and appear fragmented relative to orthologs in free-living genomes are colored red; finally, the size of each circle represents d*S*, a proxy for evolutionary divergence.

In contrast, nontransposase gene duplicates in *Sod. endolongispinus*, which comprise 101 genes, have an average d*S* of 1.3, suggesting that they likely duplicated prior to host restriction. Gene duplication prior to divergence of nontransposase genes is also supported by the fact that orthologs to most duplicated genes are also encoded as duplicates on the genome of *Sodalis* HS. Long-term maintenance of gene duplicates is considered rare in prokaryotic genomes (Hooper and Berg 2003); however, in certain bacterial species, gene duplicates do persist and can accumulate to a considerable fraction of the genome (Gevers et al. 2004; Miller et al. 2011). In contrast to *Sod. endolongispinus*, we find that *Sym. endolongispinus* does not appear to have undergone an expansion of transposases.

Many of the identified transposase genes appear to have been pseudogenized in some way. In *Sod. endolongispinus* and *Sym. endolongispinus*, 26% and 70%, respectively, of all identified transposases have been flagged as pseudogenes. The vast majority of these pseudogene predictions are based on the shorter length of each transposase relative to the closest homologs available in NCBI. These short pseudogene fragments are likely caused by deletions, which may have been preceded by some kind of nonsense mutation, or frameshift-inducing indel. There are also some transposases that appear to have acquired nonsense mutations and exist as multiple fragments on the genome. The fact that many transposases have become pseudogenized is not unique to the *P. longispinus* endosymbionts. Other *Sodalis* and *Symbiopectobacterium*-related symbionts show similar levels of pseudogenization among their transposases (supplementary file 5, Supplementary Material online).

Both Gammaproteobacterial Endosymbionts Show Complementary Patterns of Gene Pseudogenization and Loss in Amino Acid and Vitamin Biosynthesis Pathways

Although the genomes of the gammaproteobacterial symbionts of *P. longispinus* are still large, the pseudogenization of nearly half of their genes allows us to ask whether gene inactivation events show nascent signals of the interdependency that is common in more established endosymbionts (Martin and Herrmann 1998; Shigenobu 200l; Wu et al. 2006; Gosalbes et al. 2008; McCutcheon and Moran 2010; Lamelas et al. 2011; Sloan and Moran 2012; Husník et al. 2013; López-Madrigal et al. 2013; Bennett et al. 2014; Santos-Garcia et al. 2014; Luan et al. 2015; Husník and McCutcheon 2016; Szabó et al. 2017; Ankrah et al. 2020). Clear patterns of complementary gene loss and retention have been observed in other mealybug symbioses that host intra-*Tremblaya* gammaproteobacterial symbionts, but in these other cases, the gammaproteobacterial endosymbionts have highly reduced and gene-dense genomes of less than 1 Mb, consistent with much longer periods of host restriction (Husník and McCutcheon 2016; Szabó et al. 2017).

We find that *Sym. endolongispinus* and *Sod. endolongispinus* show signs of nascent complementarity in gene loss and retention. This pattern is most clear in key host-required pathways used to build essential amino acids and vitamins (fig. 4*A*). For example, the pathways for biosynthesis of the amino acids histidine, cysteine, arginine, threonine, and methionine show signs of partitioning between both gammaproteobacterial genomes through reciprocal pseudogene formation and gene loss. Some of the biosynthetic genes missing or pseudogenized from both *Sod. endolongispinus* and *Sym. endolongispinus* (e.g., *hisF*, *ribC*, *bioA*, *bioB*) are encoded either by *Tremblaya* or on the host genome as bacterial HGTs. There are also many pathway components that remain redundant in the system, with multiple gene copies present between the symbiotic partners (supplementary fig. 5, Supplementary Material online). For example, three of the genes responsible for lysine biosynthesis (*dapA*, *dapB*, and *dapF*) are encoded on the host as HGTs, but these genes are also retained in both *Sod. endolongispinus* and *Sym. endolongispinus*. There are also two instances where a required gene (*argA* [arginine] and *bioC* [biotin]) is missing completely from the symbiosis. These genes are also missing in other symbioses, and it is possible that their roles have been taken over by host proteins of eukaryotic origin (Husník et al. 2013).

Core Metabolic and Cell Structural Genes in Gammaproteobacterial Genomes Are Strongly Retained

Contrary to the pattern of complementary degradation in pathways for amino acid and vitamin biosynthesis, genes that are part of the core metabolic and cell structural pathways show strong retention in both *Sod. endolongispinus* and *Sym. endolongispinus* (figs. 4 and 5). Specifically, genes for glycolysis, pentose phosphate, and the acetate node, as well as other essential pathways (e.g., iron–sulfur cluster biosynthesis, tRNA modification), are completely intact on both of the young endosymbiont genomes in *P. longispinus* (fig. 4).

Finally, we investigated the pathway for peptidoglycan (PG) biosynthesis. PG is an important component of the bacterial cell envelope; it provides rigidity and shape to most bacterial cells (Otten et al. 2018). We have previously shown that in a related mealybug species, *Planococcus citri*, PG is produced by a biosynthetic pathway split between horizontally acquired genes encoded on the host genome and genes on the gammaproteobacterial endosymbiont genome (Bublitz et al. 2019). However, in *P. citri*, *Tremblaya* harbors an ancient and long-established gammaproteobacterial symbiont, *Ca.* Moranella endobia (hereafter, *Moranella*), which has a highly reduced genome with many deleted PG-related genes. *P. citri* and *P. longispinus* are somewhat closely related mealybugs and share the same PG-related bacterial HGTs on their nuclear genomes (Husník and McCutcheon 2016). Here, we find that the core PG biosynthesis pathway is intact in *Sym.*

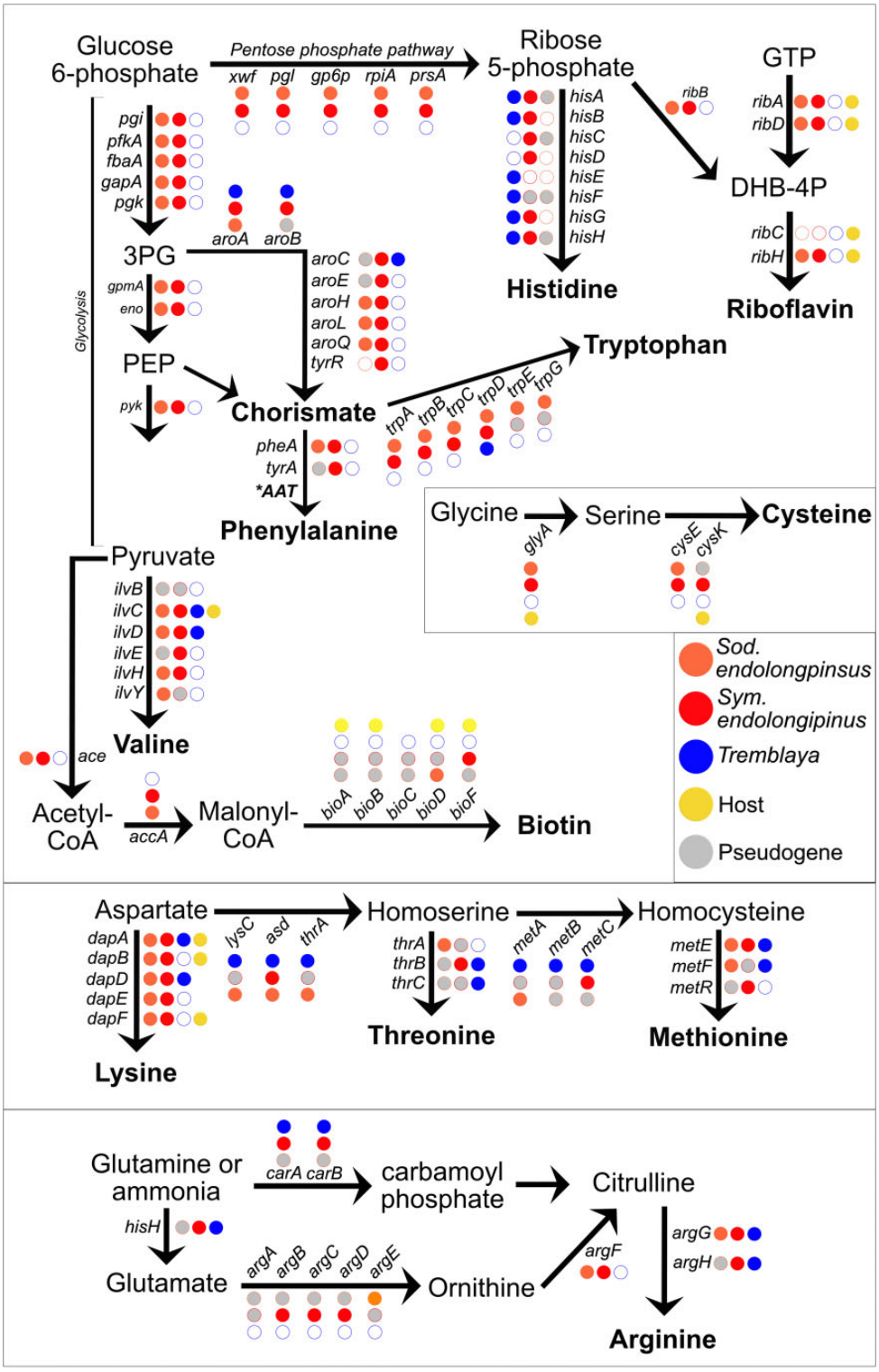

Fig. 4.—Distribution of metabolic genes in the *Pectobacterium longispinus* symbiosis. Presence, absence, and pseudogenes among the various biosynthetic pathways in *P. longispinus*. Also shown are the central metabolism pathways (glycolysis, pentose phosphate, and acetate node). Pseudogenes are colored gray. The presence of a gene on the host genome (either native or from HGT) is shown as a filled yellow circle. Essential amino acids and vitamins provided by the endosymbionts are shown in bold.

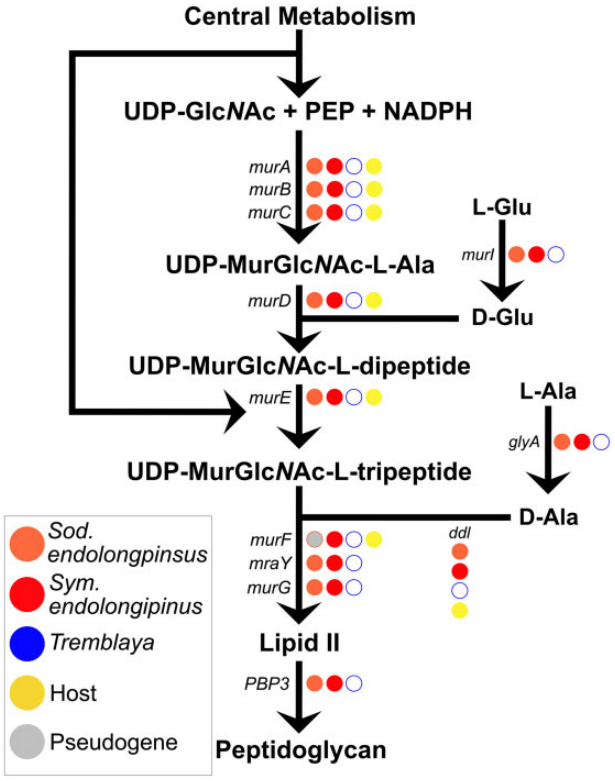

Fig. 5.—The peptidoglycan biosynthesis pathway in *Pectobacterium longispinus* symbiosis. Presence, absence, and pseudogenes within the peptidoglycan biosynthesis pathway in *P. longispinus* endosymbionts. Pseudogenes are colored gray. The presence of a gene on the host genome is shown as a filled yellow circle.

endolongispinus (fig. 5*A*). In *Sod. endolongispinus*, however, one PG-related gene, *murF*, has acquired a single-base frameshift-causing insertion in the middle of the gene (supplementary file 6, Supplementary Material online). This frameshift resulted in an internal stop-codon, causing Prodigal to predict two separate open-reading frames for this gene (supplementary file 3, Supplementary Material online). Interestingly, *murF* is present as an HGTs on the *P. longispinus* genome (Husník and McCutcheon 2016), but it is unclear if this HGT somehow complements the early loss of *murF* in *Sod. endolongispinus* in a manner similar to that in *P. citri* (Bublitz et al. 2019).

Discussion

Gammaproteobacterial Endosymbionts in P. longispinus Are of Recent Origin

We conclude that the gammaproteobacterial endosymbionts in *P. longispinus* mealybugs have been introduced into a host-restricted lifestyle relatively recently, on a timescale roughly similar (tens to hundreds of thousands of years) to other young endosymbionts of insects and nematodes (Toh et al. 2006; Burke and Moran 2011; Clayton et al. 2012; Boyd et al. 2016; Oakeson et al. 2014; Martinson et al. 2020). We base this conclusion on three features of their genomes. First, their genome sizes are large, comparable with those of free-living bacteria (table 1 and fig. 3) (Husník and McCutcheon 2016), showing that they have not yet undergone most of the genome reduction seen in more established bacterial endosymbionts (McCutcheon and Moran 2011). Second, a phylogenomic tree, consisting of endosymbionts (old and young) and free-living bacteria, shows the gammaproteobacterial endosymbionts of *P. longispinus* on short branch lengths (fig. 2), indicating that they have not yet experienced the rapid sequence evolution typical of older endosymbiotic bacteria (Moran 1996). The branch lengths of *P. longispinus* gammaproteobacteria are similar to those of other recently acquired endosymbionts and are substantially shorter than the branch lengths of older symbionts that have undergone millions of years of genome erosion. Third, their GC-contents at 4-fold degenerate sites in coding regions remain relatively high (supplementary file 4, Supplementary Material online), whereas older endosymbionts typically show pronounced AT biases at these sites (Wernegreen 2002; Van Leuven and McCutcheon 2012).

We attempted to infer which gammaproteobacterial endosymbiont might have been established first within *P. longispinus* bacteriocytes. The lower average d*S* and shorter branch length relative to its closest free-living relative (fig. 2) suggests that *Sod. endolongispinus* is the younger of the two gammaproteobacterial endosymbionts in *P. longispinus*. This is consistent with our rough estimate of 68,000 years as the divergence time between *Sod. endolongispinus* and *Sodalis* HS compared with an estimated divergence of 466,000 years for *Sym. endolongispinus* and *P. carotovorum* (Martinson et al. 2020). It is important to emphasize that these dates are extremely speculative. Additionally, it is possible that the longer branch length of *Sym. endolongispinus* (as well as the other *Symbiopectobacterium* symbionts) relative to the free-living *Pectobacterium/Brenneria* spp. is due to the fact that a close relative to the *Symbiopectobacterium* has not yet been sequenced.

Creation of Pseudogenes Is Likely Coupled with Rapid Deletion

During their brief period of host restriction and vertical transmission, *Sym. endolongispinus* and *Sod. endolongispinus* have accumulated thousands of pseudogenes (fig. 3). This is in stark contrast to nonendosymbiotic bacteria, where pseudogenes have been reported to account for only 1–5% (in some cases, as high as 8%) of the genetic repertoire (Liu et al. 2004; Lerat and Ochman 2005). The high level of gene inactivation in both *Sym. endolongispinus* and *Sod. endolongispinus* genomes is caused by many frameshift-causing indels and nonsense mutations, resulting in either truncated, run-on,

and fragmented genes (supplemental files 3 and 4, Supplementary Material online).

A small proportion of apparently functional genes appear to be cryptic pseudogenes, or genes that have elevated d*N*/d*S* values (>0.3) relative to orthologs in a closely related non-endosymbiont genome (Clayton et al. 2012; Oakeson et al. 2014; Van Leuven et al. 2014; Burke and Moran 2011). However, the vast majority of both intact and broken genes have low d*N*/d*S* values consistent with strong purifying selection. This suggests that pseudogenes in *Sym. endolongispinus* and *Sod. endolongispinus* have formed very recently and have not yet had time to accumulate substitutions that would elevate their d*N*/d*S* values. As noted in previous studies of pseudogene flux in several strains of *Salmonella* (Kuo and Ochman 2010) and consistent with the previously reported deletional bias in bacterial genomes (Mira et al. 2001; Kuo and Ochman 2009; Burke and Moran 2011), it is likely that deletion of pseudogenes happens quickly, on time scales shorter than these genes can accumulate significant numbers of nonsynonymous sequence substitutions.

Recent Transposase Expansion a Common, but Not Universal, Feature of Early Genomic Disruption in Endosymbionts

Sod. endolongispinus encodes 220 transposases, over an order of magnitude greater than its closest sequenced free-living relative. The high number of transposases in *Sod. endolongispinus* is consistent with previous reports of IS expansion as a by-product of relaxed selection on large parts of the genome as well as a mechanism for genome rearrangement and reduction (Mahillon and Chandler 1998; Plague et al. 2008; Belda et al. 2010; Schmitz-Esser et al. 2011; Oakeson et al. 2014; Hendry et al. 2018). Indeed, other recently acquired *Sodalis* endosymbionts (SOPE and *S. glossinidius*) also encode high numbers of transposases (Clayton et al. 2012; Oakeson et al. 2014). Each *Sodalis* endosymbiont encodes different types and distributions of IS families (supplementary fig. 2*A*, Supplementary Material online). Certain IS families present in the other *Sodalis* endosymbionts encode transposases that are not found in *Sodalis* HS (e.g., IS21 in SOPE, IS1 in *Sodalis* sp. SCIS). Given that ISs are highly dynamic, moving within and between genomes (Touchon and Rocha 2007), it is possible that even a very close relative to *Sodalis* HS could have radically different types and distributions of ISs.

Similar to *Sod. endolongispinus*, SOPE and *S. glossinidius* have likely undergone an IS expansion very recently, since the vast majority of the transposase sequences fall into only a handful of sequence clusters composed of nearly identical copies of just a few families per genome (supplementary file 1 and supplementary fig. 2*B*, Supplementary Material online) (Gil et al. 2008; Oakeson et al. 2014). Because transposases are known for facilitating genomic deletions (Mahillon and Chandler 1998) and have proposed to be involved in genome reduction in endosymbionts (Siguier et al. 2014), it is possible that this deletional tendency of transposases results in the elimination of IS elements themselves over time (Plague et al. 2008; Schmitz-Esser et al. 2011; Siguier et al. 2014). Consistent with this idea, the genomes of older endosymbionts encode no or very low numbers of IS elements (supplementary fig. 2, Supplementary Material online).

IS proliferation does not appear to be a universal phenomenon in young endosymbionts, at least not among the *Sodalis* and *Symbiopectobacterium* endosymbionts screened here. Several *Sodalis* endosymbionts (e.g., *Sodalis* sp. TME1, *Sodalis* sp. SCIS), whose genomes are comparable in size to *Sod. endolongispinus* (supplementary table 1, Supplementary Material online), appear to have few or no transposases (supplementary fig. 2, Supplementary Material online). None of the *Symbiopectobacterium* symbionts encodes nearly as many transposases as *Sod. endolongispinus* or *S. glossinidius* but has similar or lower numbers of transposases relative to *P. wasabiae* and *P. cartovorum*, the closest nonendosymbiotic relatives of the *Symbiopectobacterium* clade. The relative lack of transposases is surprising, given the vast amount of superfluous genome space in these young endosymbiont genomes in which transposases could insert themselves (supplementary table 1, Supplementary Material online). It is possible that there are large and diverse populations of free-living *Sodalis* and *Symbiopectobacterium* strains in nature that vary in IS content, and that the amount of IS proliferation that occurs in a newly established endosymbiont reflects the IS load of the ancestral free-living strain. Although it is also possible that the dearth of detectable transposase genes is caused by the low assembly quality of some of the endosymbiont genomes that are highly fragmented (e.g., >100 contigs) (supplementary table 2, Supplementary Material online). The highly fragmented nature of these short-read assemblies can, in part, be caused by the prevalence of identical or nearly identical transposases that cannot be resolved using short reads alone. For example, the assembly corresponding to *Sym. endolongispinus* published by Martinson et al. (2020) comprised 83 contigs (using only the Illumina reads published by Husník and McCutcheon 2016), with only 6 transposases detected using the ISfinder software that is included in the Prokka annotation pipeline. Our hybrid assembly using PacBio and Illumina reads resolved this genome into 9 contigs, from which we were able to identify 36 transposases. Similarly, our Illumina-only assembly of the *Sod. endolongispinus* genome resulted in 109 contigs, with only 7 identifiable transposases; our PacBio/Illumina hybrid assembly resolved the genome of *Sod. endolongispinus* into 3 contigs with 220 identifiable transposases. It seems likely that the estimated number of transposases in genomes

generated from short-read data alone are significantly underestimated, with near-identical ISs collapsing into small unassembled contigs.

Establishment of Interdependence Occurs Early During Endosymbiont Establishment

Previous research has demonstrated that newly evolved endosymbiotic bacteria lose genes in response to the preexisting genetic inventory of their cosymbionts and (if present) HGTs on the host genome (Wu et al. 2006; McCutcheon and Moran 2007; Nikoh and Nakabachi 2009; McCutcheon et al. 2009; Sloan and Moran 2012; Husník et al. 2013; Luan et al. 2015; Nowack et al. 2016). The occurrence of two recently acquired endosymbionts in *P. longispinus* presented us with a unique opportunity to investigate the inception of genomic complementarity and metabolic interdependence in a complex four-way symbiosis.

During this early period of host restriction, we might expect to see more rapid gene loss in pathways whose precursors, intermediates, and products are more easily transported between different members of the symbiosis. This can include metabolites for which dedicated transporters (e.g., amino acid permeases) already exist in the free-living predecessor. For example, *Sod. endolongispinus* encodes genes for the transport of histidine and biotin, possibly contributing to the rapid loss of the biotin and histidine biosynthesis pathways in that endosymbiont (fig. 4). Many genes that are part of amino acid and vitamin metabolism are either encoded on the host's nuclear genome as HGTs from bacteria or on *Tremblaya*'s diminutive genome, relieving the need for these new gammaproteobacterial symbionts to continue maintaining these genes. Consistent with this, we observe loss and pseudogenization of many pathway components for amino acid and vitamin biosynthesis in *Sym. endolongispinus* and *Sod. endolongispinus* (fig. 4*A*), suggesting that genes in these pathways are lost rapidly and in response to genes already present in the symbiosis.

As *Sym. endolongispinus* and *Sod. endolongispinus* are relatively new to a host-dependent lifestyle, they still encode many genes redundant with other genes present in the system (fig. 4*B*). Most of these redundant genes appear to be undergoing strong purifying selection, evident from their low d*N*/d*S* values (supplementary files 3 and 4, Supplementary Material online). Because other older and longer-established mealybug symbioses show little evidence of genetic redundancy across genomes (Husník and McCutcheon 2016), we suspect that many of the redundant genes in the gammaproteobacterial endosymbionts simply have not had a chance to accumulate substitutions that would break genes, elevate their d*N*/d*S* values, or delete them completely from one genome or the other.

Loss of Core Metabolic and Structural Genes Occurs More Slowly

In contrast to the rapid gene loss in pathways for amino acid and vitamin biosynthesis, genes in pathways for peptidoglycan biosynthesis and central metabolism are more strongly conserved in both of the gammaproteobacterial symbionts of *P. longispinus*. We have previously demonstrated that in *Moranella*, the ancient gammaproteobacterial symbiont of *P. citri*, peptidoglycan biosynthesis occurs in concert with bacterial genes encoded on its host's nuclear genome as HGTs (Bublitz et al. 2019). Consequently, *Moranella* has lost much of its peptidoglycan biosynthesis pathway and is presumably reliant on the import of host-derived proteins. Although this level of cellular integration represents a potential future state for *Sym. endolongispinus* and *Sod. endolongispinus*, it appears that this level of integration has not yet been achieved in the relatively short amount of time that these symbionts have been inhabiting *P. longispinus*.

We hypothesize that the generally stronger retention of the peptidoglycan biosynthesis pathway throughout the early stages of endosymbiosis is due to the difficulty of integrating pathways that require the shuttling of complex molecules (such as PG precursors) or proteins between different cellular compartments. The loss of a key gene of the peptidoglycan biosynthesis pathway (*murF*) in *Sod. endolongispinus* is therefore quite interesting given that the protein product of this gene has been shown to be imported by the gammaproteobacterial endosymbiont in a related mealybug (Bublitz et al. 2019). It is possible that the repeated recruitment and maintenance of endosymbionts from the *Sodalis* genus (Husník and McCutcheon 2016) has made mealybugs particularly suited for rapid cellular integration of *Sodalis* relatives after infection. Rapid *Sodalis* adaptation to mealybug endosymbiosis is consistent with stronger conservation of peptidoglycan biosynthesis in *Sym. endolongispinus*, even though we estimate that *Sym. endolongispinus* is the older of the two gammaproteobacterial endosymbionts. More rapid *Sodalis* adaptation is also supported by the patterns observed in the degradation of pathways for the synthesis of essential amino acids and vitamins: out of those pathway components that are encoded by *Tremblaya* or the host genome as HGTs, 21 genes are lost by *Sod. endolongispinus*, whereas only 11 are lost by *Sym. endolongispinus* (fig. 4). Members of the *Symbiopectobacterium* clade do not seem to commonly infect mealybugs, as only one of seven mealybug species for which we have genomic data houses a *Symbiopectobacterium*-related symbiont (Husník and McCutcheon 2016; Szabó et al. 2017). A possible preference of mealybugs toward recruitment of *Sodalis*-related endosymbionts may be due to a combination of factors, including as of yet unknown factors within the host, as well as the genetic repertoire of the infecting bacteria, although it just may be that mealybugs interact more frequently with *Sodalis* in nature.

Materials and Methods

Insect Rearing

Pseudococcus longispinus populations were reared on sprouted potatoes (fig. 1*A*) at 25 °C, 77% relative humidity, and a 12 h light/dark cycle in a Percival 136LL incubator.

RNA-Fluorescence In Situ Hybridization

Whole *P. longispinus* individuals of the second and third instar developmental stage were submerged in Ringer solution (3 mM CaCl2 × $2H_2O$, 182 mM KCl, 46 mM NaCl, 10 mM Tris base; adjusted to pH 7.2) and carefully opened for better buffer infiltration. Samples were transferred into Carnoy's fixative (EtOH:chloroform:acetic acid; 6:3:1) and fixed overnight at 4 °C. Tissue samples were then dehydrated in a graded ethanol series from 70% to 100% ethanol. Samples were transferred into tissue bags and cassettes for paraffin embedding using a Leica ASP 300 Tissue Processor. Ethanol was exchanged for methyl salicylate and then incubated in 100% xylene before infiltration with paraffin. Each individual sample was embedded in a single paraffin block, semi-thin sections (5–6 μm) were prepared with a microtome and mounted onto microscopy slides.

Sections designated for RNA-FISH experiments were deparaffinized in xylene and rehydrated in a graded ethanol series (100–30% ethanol). Tissue sections were then prehybridized in hybridization buffer (900 mM NaCl, 20 mM Tris–HCl pH 7.5, 35% formamide). Hybridization was performed by adding 1.5–2 μl of the probe targeting the *Pectobacterium*-related endosymbiont (5′[Cy3]-ccacgcctcaagggcacaacctc; 100 μM) to each 100 μl hybridization buffer and incubated at 40 °C. Samples were then briefly rinsed in wash buffer (70 mM NaCl, 20 mM Tris–HCl pH 7.5, 5 mM EDTA pH 8.0, 0.01% SDS) before mounting the slides with hybridization buffer supplemented with 1.5–2 μL of each probe (per 100 μl buffer) targeting the *Sodalis*-related endosymbiont (5′[Cy5]-aaagccacggctcaaggccacaacctt; 100 μM) and *Tremblaya* (5′[fluorescein]-gccttagcccgtgctgccgtac; 100 μM), followed by overnight incubation at 30 °C. Slides were then washed in washing buffer at 30 °C and counterstained with Hoechst in washing buffer. After another washing step, sample slides were rinsed in dH_2O, mounted with FluorSave™ Reagent (Sigma Millipore) and analyzed by confocal laser scanning microscopy with a Zeiss LSM 880. Images were processed using Fiji version 1.0.

Electron Microscopy

Bacteriomes were dissected from second and third instar *P. longispinus* individuals as previously described (Bublitz et al. 2019). Isolated bacteriomes were prefixed with 3% glutaraldehyde, 1% paraformaldehyde, and 5% sucrose in 0.1 M sodium cacodylate trihydrate for 12–24 h at 4 °C, then rinsed briefly with cacodylate buffer. Bacteriomes were placed into brass planchettes (Ted Pella Inc.) prefilled with cacodylate buffer + 10% 70 kD Ficoll (extracellular cryoprotectant; Sigma) and ultra-rapidly frozen with a HPM010 High Pressure Freezing machine (BalTec/ABRA, Switzerland). Vitreously frozen samples were transferred, under liquid nitrogen, to Nunc cryovials (Thermo-Fisher Scientific) containing 2% OsO_4 and 0.05% uranyl acetate in acetone and placed into an AFS-2 Freeze Substitution Machine (Leica Microsystems, Austria). Samples were freeze-substituted at −90 °C for 72 h, warmed to −20 °C over 12 h, held at −20 °C for an additional 12 h, and then brought to room temperature. Samples were rinsed 4× with acetone, infiltrated into Epon-Araldite resin (Electron Microscopy Sciences, Port Washington, PA), then flat-embedded between two Teflon-coated glass microscope slides. Resin was polymerized at 60 °C for 24–48 h. Embedded samples were observed by phase-contrast microscopy to ascertain specimen quality and to select appropriate regions for electron microscopy (EM). Blocks of tissue (typically containing a single bacteriome) were excised with a scalpel and glued to plastic sectioning stubs. Serial semi-thick (150–300 nm) sections were cut with a UC6 ultramicrotome (Leica Microsystems) using a diamond knife (Diatome Ltd, Switzerland) and collected onto Formvar-coated copper-rhodium 1 mm slot grids (Electron Microscopy Sciences). Grids were stained with 3% uranyl acetate and lead citrate; then, 10 nm colloidal gold particles were applied to both sides of the sections to serve as fiducial markers for subsequent tomographic image alignment.

Dual-Axis Tomography

Grids were placed in a dual-axis tomography specimen holder (Model 2040; E.A. Fischione Instruments Inc., Export PA) and viewed with a Tecnai TF-30ST TEM at 300 KeV. Dual-axis tilt-series were acquired automatically using the SerialEM software package (Mastronarde 2005) and recorded digitally with a 2k × 2k CCD camera (XP1000; Gatan Inc., Pleasanton, CA). Briefly, sections were tilted ±64° with images taken at 1° increments. The grid was then rotated 90° and a similar tilt-series was recorded around the orthogonal axis. Tomograms were calculated, joined, and analyzed using the IMOD software package (Mastronarde 2008; Mastronarde and Held 2017) on Mac Pro and iMac Pro computers (Apple Inc.).

Sequencing and Assembly

Raw reads (Illumina HiSeq 2000) published in 2016 by Husník and McCutcheon (BioProject: PRJEB12068) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), using the SRA Toolkit v2.10.8 (SRA Toolkit Development Team). Reads were trimmed using Trimmomatic v0.36 (minimum length = 36 bp, sliding window = 4 bp, minimum quality score = 15

[ILLUMINACLIP: TruSeq3-PE : 2:30:10 LEADING : 3 TRAILING : 3 SLIDINGWINDOW : 4:15 MINLEN : 36]) (Bolger et al. 2014). For PacBio sequencing, genomic DNA was prepared from pooled mealybugs of the second and third instar stage using Qiagen Genomic Tip 500 g extraction kits, size selected for fragments >20 kb using a BluePippen device, followed by library preparation using a single molecule realtime (SMRTbell) Template Prep Kit v1.0. The resulting libraries were sequenced on 28 SMRT PacBio cells using P6 version 2 chemistry and reagents by Sci-Life labs in Uppsala, Sweden. This sequencing effort resulted in 6,101,355 reads with an average length of 9,805 bases, for a total of 59,828,022,374 bases. These reads were error corrected and trimmed, resulting in 5.05 million reads with an average sequence length of 9,318 bases. These corrected and trimmed reads were then assembled using Canu v1.6 (correctedErrorRate = 0.45, genomesize = 284 m) (Koren et al. 2017), which produced 3,049 contigs spanning 438,113,873 bases. Preliminary gammaproteobacterial contigs were extracted from the Canu assembly using the SprayNPray software (https://github.com/Arkadiy-Garber/SprayNPray; last accessed June 11, 2021). Briefly, SprayNPray uses Prodigal v2.6.3 (Hyatt et al. 2010) to predict genes from coding sequences and then queries the protein translation for each gene against NCBI's RefSeq database (release 200; Pruitt et al. 2007) using DIAMOND v2.0.4.142 (e-value cutoff 1E−6; Buchfink et al. 2015). Putative endosymbiont contigs were then extracted from the larger assembly based on gene density, GC-content, and taxonomy of top DIAMOND hits (to *Sodalis*- and *Pectobacterium/Brenneria*-related spp.) to each contig. These contigs were then used to identify and extract all Illumina and PacBio reads associated with the gammaproteobacterial symbionts. Identification of endosymbiont-derived Illumina short reads was performed using Bowtie2 v2.3.4.1 (Langmead and Salzberg 2012). Identification of endosymbiont-derived PacBio reads was performed using BLASR v5.1 (Chaisson and Tesler 2012). About 3.2% of all PacBio reads mapped to the CANU-assembled contigs affiliated with the gammaproteobacterial endosymbionts. Of the 124.5 million Illumina read pairs, 4.8% mapped to the crude gammaproteobacterial endosymbiont contigs.

Once these gammaproteobacterial subsets of short and long reads were identified and extracted, Unicycler v0.4.8 (Wick et al. 2017) was used, with "normal" mode (minimum bridge quality = 10) to carry out a hybrid SPAdes v3.13.0 (Bankevich et al. 2012) assembly (default k-mers). This resulted in 15 contigs, out of which the two gammaproteobacterial genomes were binned using a combination of metrics, including coverage of Illumina reads mapped against the final assemblies. Coverage of the short Illumina reads was estimated using the *jgi_summarize_bam_contig_depths* script from the MetaBAT package (Kang et al. 2019). Since the closest phylogenomic affiliations of each gammaproteobacterial symbiont are known (Husník and McCutcheon 2016), we also used BlastP (Camacho et al. 2009) to compare predicted proteins from each contig against NCBI's RefSeq database (Pruitt et al. 2007). Proteins were derived from Prodigal's gene predictions (Hyatt et al. 2010), and the phylogenetic affiliation of each contigs' proteins was inferred by the top BlastP hits from NCBI's RefSeq database (Pruitt et al. 2007).

Phylogenomic Analysis

Phylogenomic analysis was carried out using GToTree v1.5.38 (Lee 2019) and RAxML (Stamatakis 2014). Briefly, single-copy genes were identified using a set of HMMs for genes common to gammaproteobacteria (Lee 2019). As part of the GToTree pipeline, single-copy genes are identified using HMMER v3.2.1 (Johnson et al. 2010), aligned with Muscle v3.8 (Edgar 2004), and concatenated. We then used these concatenated alignments to build a phylogenomic tree with RAxML, with 100 bootstraps (-N 100), the PROTCAT model for amino acid substitution, and the BLOSUM 62 amino acid matrix (-m PROTCATBLOSUM62) (Stamatakis 2014). Phylogenomic trees were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/; last accessed June 11, 2021).

Average nucleotide and amino acid identities between endosymbionts were estimated using a Blast-based approach: genes from each endosymbiont were queried in a Blast search against all other symbionts, requiring an e-value of 1E−10 or lower, and allowing for the reporting of only one top Blast hit. We then calculated the average sequence identity among all of the Blast matches between each endosymbiont comparison.

Annotation and Biosynthetic Pathway Reconstruction

Each endosymbiont was annotated using Prokka v1.14.6 (Seemann 2014). As part of Prokka's pipeline, coding regions are detected using Prodigal; noncoding RNA sequences were also identified: tRNAs and tmRNAs using Aragorn (Laslett and Canback 2004) and rRNAs using RNAmmer (Lagesen et al. 2007). Prokka annotation also included identification of transposases, using the ISfinder database of insertion sequences (Siguier et al. 2006). Genes were also annotated using the GhostKOALA v.2.2 web server, which uses the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology database (Kanehisa et al. 2016). Biosynthetic pathways for amino acids, vitamins, peptidoglycan, and translation-related genes were manually identified and reconstructed from these annotations and organized into pathways. HGTs present on the mealybug genome were previously identified (Husník and McCutcheon 2016; Bublitz et al. 2019) and further confirmed with the SprayNPray software.

Pseudogene Prediction

Candidate pseudogenes were identified using the Pseudofinder software v1.0 (https://github.com/filip-husnik/pseudofinder; last accessed June 11, 2021), with DIAMOND v2.0.4.142 (Buchfink et al. 2015) as the search engine (–diamond) to find each gene's closest homologs in NCBI's RefSeq (Pruitt et al. 2007) database. This allowed us to identify pseudogenes based on length and gene fragmentation due to early stop codons. Proteins shorter than 75% (–length_pseudo 75) compared with the average length of the 15 top homologs (–hitcap 15, –evalue 1E−4) from RefSeq were flagged as potential pseudogenes. Additionally, genes with internal stop codons and frameshift mutations were also flagged as pseudogenes. These fragmented genes were identified by Pseudofinder by finding adjacent proteins that have the same protein sequence as their top DIAMOND hit. For adjacent genes to be considered as fragmented parts of the same ancestral gene, we used a distance cutoff of 2,000 bp (–distance 2,000). The length of each gene and pairwise homology (between the putative fragments) was also taken into consideration to exclude adjacent genes that represent gene duplication events, which may have resulted in tandem-encoded duplicate genes. Nongenic regions, in which Prodigal did not detect any genes, were also compared against NCBI's RefSeq database, to identify pseudogenized genes that were missed by Prodigal's gene-finding algorithm. Nongenic regions required at least five DIAMOND matches to proteins in the RefSeq database (–intergenic_threshold 0.3) to be considered as pseudogenes.

Using DIAMOND BlastP, Pseudofinder also compared genes from each endosymbiont to genes encoded by its closest free-living relative, inferred from phylogenomic analysis and average amino acid identity. We identified *P. wasabiae* as the closest free-living relative to one endosymbiont and *Sodalis praecaptivus* HS as the closest relative to the other. Using PAL2NAL v14 (Suyama et al. 2006), Pseudofinder generates codon alignments for each ortholog pair, then, using Codeml v4.9j (Yang 2007), calculates d*N*/d*S* values for each pairwise comparison. We provide the control file (codeml.ctl) containing the parameters used by Codeml in the following GitHub repository: https://github.com/Arkadiy-Garber/PLON-genome-paper; last accessed June 11, 2021. We required d*S* to be greater than 0.001 and lower than three for d*N*/d*S* calculation (-m 0.001, -M 3). This allowed us to infer cryptic pseudogenes, or genes that are likely undergoing relaxed selection but have not acquired any obvious inactivating mutations (Clayton et al. 2012; Oakeson et al. 2014; Van Leuven et al. 2014). We used a d*N*/d*S* cutoff of 0.3 (Oakeson et al. 2014), flagging genes as pseudogenes if their d*N*/d*S* values are higher than this threshold (–max_dnds 0.3).

Pseudogene calls from Pseudofinder were manually inspected, using AliView (Larsson 2014) to confirm gene fragmentation, d*N*/d*S* values, and other inactivating mutations.

Identification of Duplicated Genes

We used ParaHunter to identify gene duplicates in the endosymbiont genomes (Miller et al. 2020; https://github.com/Arkadiy-Garber/ParaHunter; last accessed June 10, 2021). This software uses MMseqs2 v12.113e3 (Steinegger and Söding 2017) to identify homologous gene clusters within each genome. For this analysis, we used a cutoff of 50% amino acid identify (-m 0.5) over at least 50% of the length (-l 0.5) of target sequence. After clusters are identified, within-cluster analysis is carried out, where pairwise amino acid and nucleotide alignments are converted to codon alignments using PAL2NAL (Suyama et al. 2006), and then d*N*/d*S* is calculated using Codeml (Yang 2007). Calculation of d*N*/d*S* was only performed on those gene pairs where d*S* values were greater than 0.001 and lower than 3.

Additional Scripts and Plotting

Additional custom python scripts were used to process the data presented in this study. These scripts are all annotated and available in the following GitHub repository: https://github.com/Arkadiy-Garber/PLON-genome-paper; last accessed June 11, 2021. Many plots presented in this study were made in R (R Core Team 2013), using the following packages: ggplot2 (Wickham 2009) and reshape (Wickham 2007).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Diane Brooks, Paul Caccamo, Filip Husník, Genevieve Krause, Piotr Łukasik, Mitch Syberg-Olsen, Dan Vanderpool, Catherine Armbruster, and Travis Wheeler for helpful discussions and insights during the course of this project. We thank the Caltech Kavli Nanoscience Institute for maintenance of the TF-30 electron microscope. This work was supported by the National Science Foundation (IOS-1553529 to J.P.M.), the Gordon and Betty Moore Foundation (GBMF5602 to J.P.M.), the National Aeronautics and Space Administration Astrobiology Institute (NNA15BB04A to J.P.M.), and the National Institute of Allergy and Infectious Diseases (2 P50 AI150464 to P.J.B.).

Data Availability

PacBio reads were deposited to the SRA, under NCBI BioProject PRJNA719553. Genome assemblies for the gammaproteobacterial endosymbionts are available under NCBI BioProject PRJEB12068 (BioSamples SAMN17910461 for the *Sodalis*-related endosymbiont and SAMN17910462 for the

Symbiopectobacterium-related symbiont). Genome sequences and annotation data for the two gammaproteobacterial endosymbionts were also made available via figshare: Genome sequence and Prokka-annotation are available at https://doi.org/10.6084/m9.figshare.13632407.v1 for the *Pectobacterium*-related symbiont and https://doi.org/10.6084/m9.figshare.13632398.v2 for the *Sodalis*-related symbiont. Pseudogene predictions are available at https://doi.org/10.6084/m9.figshare.13632419.v1 for the *Pectobacterium*-related symbiont and https://doi.org/10.6084/m9.figshare.13632416.v1 for the *Sodalis*-related symbiont. Files used in the analysis of other *Sodalis*- and *Symbiopectobacterium*-related endosymbionts are available here: https://doi.org/10.6084/m9.figshare.13661189.

Literature Cited

Ankrah NYD, et al. 2020. Syntrophic splitting of central carbon metabolism in host cells bearing functionally different symbiotic bacteria. ISME J. 14(8):1982–1993.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Baumann L, et al. 2002. The genetic properties of the primary endosymbionts of mealybugs differ from those of other endosymbionts of plant sap-sucking insects. Appl Environ Microbiol. 68(7):3198–3205.

Baumann P. 2005. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol. 59:155–189.

Belda E, et al. 2010. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. BMC Genomics 11:449.

Bennett GM, et al. 2014. Differential genome evolution between companion symbionts in an insect-bacterial symbiosis. MBio. 5(5):e01697–14.

Bolger AM, et al. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Boyd BM, et al. 2016. Two bacterial genera, *Sodalis* and *Rickettsia*, associated with the seal louse *Proechinophthirus fluctus* (Phthiraptera: Anoplura). Appl Environ Microbiol. 82(11):3185–3197.

Bublitz DC, et al. 2019. Peptidoglycan production by an insect-bacterial mosaic. Cell 179(3):703–712.e7.

Buchfink B, et al. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Buchner P. 1965. Endosymbiosis of animals with plant microorganisms. New York: John Wiley & Sons.

Burke GR, Moran NA. 2011. Massive genomic decay in Serratia symbiotica, a recently evolved symbiont of aphids. Genome Biol Evol. 3:195–208.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238.

Chari A, et al. 2015. Phenotypic characterization of *Sodalis praecaptivus* sp. nov., a close non-insect-associated member of the *Sodalis*-allied lineage of insect endosymbionts. Int J Syst Evol Microbiol. 65(Pt 5):1400–1405.

Clayton AL, et al. 2012. A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. PLoS Genet. 8(11):e1002990.

Douglas AE. 2006. Phloem-sap feeding by animals: problems and solutions. J Exp Bot. 57(4):747–754.

Downie DA, Gullan PJ. 2005. Phylogenetic congruence of mealybugs and their primary endosymbionts. J Evol Biol. 18(2):315–324.

D'Souza G, Kost C. 2016. Experimental evolution of metabolic dependency in bacteria. PLoS Genet. 12(11):e1006364.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Fukatsu T, Nikoh N. 1998. Two intracellular symbiotic bacteria from the mulberry psyllid *Anomoneura mori* (Insecta, Homoptera). Appl Environ Microbiol. 64(10):3599–3606.

Gardan L, Gouy C, Christen R, Samson R. 2003. Elevation of three subspecies of *Pectobacterium carotovorum* to species level: *Pectobacterium atrosepticum* sp. nov., *Pectobacterium betavasculorum* sp. nov. and *Pectobacterium wasabiae* sp. nov. Int J Syst Evol Microbiol. 53(Pt 2):381–391.

Gatehouse LN, Sutherland P, Forgie SA, Kaji R, Christeller JT. 2012. Molecular and histological characterization of primary (betaproteobacteria) and secondary (gammaproteobacteria) endosymbionts of three mealybug species. Appl Environ Microbiol. 78(4):1187–1197.

Gevers D, Vandepoele K, Simillon C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12(4):148–154.

Gil R, et al. 2008. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*. Int Microbiol. 11(1):41–48.

Gil R, et al. 2018. Tremblaya phenacola PPER: an evolutionary beta-gammaproteobacterium collage. ISME J. 12(1):124–135.

Gosalbes MJ, et al. 2008. The striking case of tryptophan provision in the cedar aphid *Cinara cedri*. J Bacteriol. 190(17):6026–6029.

Hall RJ, et al. 2020. Simulating the evolutionary trajectories of metabolic pathways for insect symbionts in the genus *Sodalis*. Microb Genom. 6:mgen000378.

Hendry TA, et al. 2018. Ongoing transposon-mediated genome reduction in the luminous bacterial symbionts of deep-sea ceratioid anglerfishes. MBio. 9(3):e01033.

Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol. 20(6):945–954.

Husník F, McCutcheon JP. 2016. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. Proc Natl Acad Sci USA. 113(37):E5416–E5424.

Husník F, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153:1567–1578.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

Johnson LS, et al. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

Kanehisa M, et al. 2016. BlastKOALA and GhostKOALA: KEGG Tools for functional characterization of genome and metagenome sequences. J Mol Biol. 428(4):726–731.

Kang DD, et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. Peer J. 7:e7359.

Koga R, Nikoh N, Matsuura Y, Meng X-Y, Fukatsu T. 2013. Mealybugs with distinct endosymbiotic systems living on the same host plant. FEMS Microbiol Ecol. 83(1):93–100.

Komaki K, Ishikawa H. 1999. Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. J Mol Evol. 48(6):717–722.

Kono M, et al. 2008. Infection dynamics of coexisting beta- and gammaproteobacteria in the nested endosymbiotic system of mealybugs. Appl Environ Microbiol. 74(13):4175–4184.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.
Kuo C-H, Ochman H. 2009. Deletional bias across the three domains of life. Genome Biol Evol. 1:145–152.
Kuo C-H, Ochman H. 2010. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 6(8):e1001050.
Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.
Lamelas A, et al. 2011. Serratia symbiotica from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. PLoS Genet. 7(11):e1002357.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30(22):3276–3278.
Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32(1):11–16.
Liu Y, et al. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. 5(9):R64.
Lee MD. 2019. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics 35(20):4162–4164.
Lerat E, Ochman H. 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. 33(10):3125–3132.
López-Madrigal S, et al. 2013. Mealybugs nested endosymbiosis: going into the 'matryoshka' system in *Planococcus citri* in depth. BMC Microbiol. 13:74.
Luan J-B, et al. 2015. Metabolic coevolution in the bacterial symbiosis of whiteflies and related plant sap-feeding insects. Genome Biol Evol. 7(9):2635–2647.
Łukasik P, et al. 2018. Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas. Proc Natl Acad Sci USA. 115(2):E226–E235.
Mahillon J, Chandler M. 1998. Insertion sequences. Microbiol Mol Biol Rev. 62(3):725–774.
Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118(1):9–17.
Martinson VG, et al. 2020. Multiple origins of obligate nematode and insect symbionts by a clade of bacteria closely related to plant pathogens. Proc Natl Acad Sci USA. 117(50):31979–31986.
Mastronarde DN. 2005. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol. 152(1):36–51.
Mastronarde DN. 2008. Correction for non-perpendicularity of beam and tilt axis in tomographic reconstructions with the IMOD package. J Microsc. 230(Pt 2):212–217.
Mastronarde DN, Held SR. 2017. Automated tilt series alignment and tomographic reconstruction in IMOD. J Struct Biol. 197(2):102–113.
McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci USA. 104(49):19392–19397.
McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol. 2:708–718.
McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10(1):13–26.
McCutcheon JP, von Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol. 21(16):1366–1372.
McCutcheon JP. et al. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A. 106(36):15394–15399.
McCutcheon JP, et al. 2019. The life of an insect endosymbiont from the cradle to the grave. Curr Biol. 29(11):R485–R495.
Miller SR, et al. 2011. Dynamics of gene duplication in the genomes of chlorophyll d-producing cyanobacteria: implications for the ecological niche. Genome Biol Evol. 3:601–613.
Miller SR, et al. 2020. Bacterial adaptation by a transposition burst of an invading IS element. bioRxiv. Available from: https://doi.org/10.1101/2020.12.10.420380.
Mira A, et al. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17(10):589–596.
Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci USA. 93(7):2873–2878.
Moran NA, et al. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 42:165–190.
Nikoh N, Nakabachi A. 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. 7:12.
Nowack ECM, et al. 2016. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. Proc Natl Acad Sci USA. 113(43):12214–12219.
Oakeson KF, et al. 2014. Genome de0generation and adaptation in a nascent stage of symbiosis. Genome Biol Evol. 6(1):76–93.
Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. Science 311(5768):1730–1733.
Otten C, et al. 2018. Peptidoglycan in obligate intracellular bacteria. Mol Microbiol. 107(2):142–163.
Parkinson JF, et al. 2017. The more, the merrier? Obligate symbiont density changes over time under controlled environmental conditions, yet holds no clear fitness consequences. Physiol Entomol. 42(2):163–172.
Pasanen M, et al. 2013. Characterization of *Pectobacterium wasabiae* and *Pectobacterium carotovorum* sub sp. *carotovorum* isolates from diseased potato plants in Finland: *Pectobacterium* strains. Ann Appl Biol. 163(3):403–419.
Plague GR, et al. 2008. Extensive proliferation of transposable elements in heritable bacterial symbionts. J Bacteriol. 190(2):777–779.
Pruitt KD, et al. 2007. NCBI reference sequences (RefSeq): a curated nonredundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(Database Issue):D61–D65.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Core Team.
Rosenblueth M, et al. 2012. Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea). J Evol Biol. 25(11):2357–2368.
Santos-Garcia D, et al. 2014. Small but powerful, the primary endosymbiont of moss bugs, *Candidatus Evansia muelleri*, holds a reduced genome with large biosynthetic capabilities. Genome Biol Evol. 6(7):1875–1893.
Schmitz-Esser S, et al. 2011. A bacterial genome in transition: an exceptional enrichment of IS elements but lack of evidence for recent transposition in the symbiont *Amoebophilus asiaticus*. BMC Evol Biol. 11:270.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.
Shigenobu S, et al. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407(6800):81–86.
Siguier P, et al. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34(Database Issue):D32–D36.
Siguier P, et al. 2014. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 38(5):865–891.
Sloan DB, Moran NA. 2012. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. Mol Biol Evol. 29(12):3781–3792.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35(11):1026–1028.
Suyama M, et al. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server Issue):W609–W612.
Szabó G, et al. 2017. Convergent patterns in the evolution of mealybug symbioses involving different intrabacterial symbionts. ISME J. 11(3):715–726.
Thao ML, et al. 2002. Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. Appl Environ Microbiol. 68(7):3190–3197.
Toh H, et al. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. Genome Res. 16(2):149–156.
Touchon M, Rocha EPC. 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol. 24(4):969–981.
Toyofuku M, et al. 2019. Types and origins of bacterial membrane vesicles. Nat Rev Microbiol. 17(1):13–24.
van Ham RCHJ, et al. 2003. Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci USA. 100(2):581–586.
Van Leuven JT, McCutcheon JP. 2012. An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. Genome Biol Evol. 4(1):24–27.
Van Leuven JT, et al. 2014. Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. Cell 158:1270–1280.
von Dohlen CD, Kohler S, Alsop ST, McManus WR. 2001. Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. Nature 412:433–436.
Wernegreen JJ. 2002. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet. 3(11):850–861.
Wick RR, et al. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 13(6):e1005595.
Wickham H. 2007. Reshaping data with the reshape package. J Stat Softw. 21:1–20.
Wickham H. 2009. Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.
Woyke T, et al. 2010. One bacterial cell, one complete genome. PLoS One 5(4):e10314.
Wu D, et al. 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. 4(6):e188.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8):1586–1591.
Yuan KX, et al. 2014. Draft genome sequence of *Pectobacterium wasabiae* strain CFIA1002. Genome Announc. 2:e00214-14.

Associate editor: Emmanuelle Lerat